Analysis and modelling of initial delay time and its impact on propagation delay of CMOS logic gates

Y.-H. Yang
C.-Y. Wu

Indexing terms: Metal-Oxide-semiconductor structures, Modelling

Abstract: The initial delay times due to the capacitive feedthrough effects in CMOS inverters are characterised and investigated. Based on the MOSFET large-signal model, the initial delay is modelled for a chain of CMOS inverters under step and ramp inputs. Optimal design that results in the minimum initial delay is obtained. Correlation between the initial delay and the propagation delay is constructed in the case of characteristic waveforms. The initial delays are found to determine the propagation delay. Applying the model to evaluate the speed performance of a scaled-down CMOS, the delay improvements for various scaling laws are compared. It is found that the most effective law in reducing the initial delay for internal circuits is the constant voltage law, whereas that for the input stage is the constant electric field law. Comparisons to SPICE simulation results are also given and good agreement is achieved.

List of symbols

$A_{s(d)}$ = area of source (drain) region
C_J = zero-bias bulk-junction bottom capacitance per unit area
C_{JSW} = zero-bias bulk-junction sidewall capacitance per unit length
$C_{bd(s)}$ = bulk-drain (source) junction capacitance
$C_{gs(d)o}$ = gate-source (drain) overlap capacitance
C_L = loading capacitance of chain of CMOS inverters in each stage
$C_{on(p)}$ = oxide capacitance of n-(p-) channel MOSFET
$DELTA$ = parameter of narrow width effect for threshold voltage
$L_{n(p)}$ = effective channel length of n-(p-) channel MOSFET
L_{eff} = effective channel length of MOSFET
$M_{Jn(p)}$ = bulk-junction bottom grading coefficient of n-(p-) channel MOSFET
$M_{JSWn(p)}$ = bulk-junction sidewall grading coefficient of n-(p-) channel MOSFET
N_{sub} = substrate doping concentration
PB = bulk-junction potential
$P_{s(d)}$ = perimeter of source (drain) region
q = magnitude of electronic charge
T_{ox} = gate oxide thickness
V_{max} = maximum drift velocity of carrier
V_{To} = threshold voltage under zero bias
$V_{Tn(p)}$ = threshold voltage of n-(p-) channel MOSFET
V_{bs} = voltage drop between bulk and source regions
V_t = thermal voltage
$W_{n(p)}$ = effective channel width of n-(p-) channel MOSFET
W_{ex} = dimension of source/drain regions in channel length direction
X_J = metallurgical junction depth
$\alpha_{s(d)}$ = geometrical factor of source (drain) junction with short-channel effect
$\beta_{n(p)}$ = transconductance parameter of n-(p-) channel MOSFET
$\lambda_{n(p)}$ = channel-length modulation parameter of n-(p-) channel MOSFET
$\gamma_{n(p)}$ = bulk threshold parameter of long channel n-(p-) channel MOSFET
$\varepsilon_{si(ox)}$ = permittivity of silicon (silicon dioxide)
μ_o = surface mobility
$\mu_{n(p)}$ = mobility of n-(p-) channel MOSFET
μ_{crit} = critical field coefficient for mobility degradation
μ_{exp} = critical field exponent for mobility degradation
μ_{tra} = transverse field coefficient
ϕ_F = Fermi potential

1 Introduction

The advantages of low-power consumption and high-noise immunity make CMOS the main technology in VLSI and ULSI [1]. Ideally, the signal of a CMOS logic gate is rail-to-rail between the power supply (V_{DD}) and ground (GND). However, transient simulations show that the actual transient voltage level at the output node of a CMOS logic gate is greater than V_{DD} or smaller than GND during certain periods. When this voltage overshoot or undershoot occurs, a charging or discharging current, opposite in direction to the normal output device currents, is generated on the device capacitance and fed to the output node of the gate to increase the voltage above V_{DD} or decrease it below GND. Owing to such a capacitive feedthrough effect, extra delay time, called the initial delay time, is needed to remove those excess feedthrough charges at the output node and recover the signal to its normal level. It is found that this initial delay time usually dominates the total signal delay [2], especially in the case that the input waveform has finite rise or fall time. The speed performance of a CMOS gate is therefore strongly affected by the initial delay times. However, it has not yet been well characterised in recent timing papers [2–7].

Paper 6768G (E10), first received 26th September 1988 and in revised form 30th January 1989

The authors are with the Department of Electronics Engineering and Institute of Electronics, National Chiao Tung University, Hsin Chu, Taiwan, Republic of China

To obtain a better understanding of initial delays, and since the behaviour of other CMOS inverting logic gates can be understood from that of the basic inverter, initial delays of CMOS inverters are investigated and analysed in this paper. Taking the physical nature of transient waveforms into consideration, and using the analytical current equations [8–11] which consider the small-geometry effects, the initial delays of CMOS inverters under step and ramp input excitations are analytically modelled. Based on the developed analytical models, the dependence of initial delay on the input slope, device parameters and the loading capacitance is investigated. Optimal device dimension resulting in minimum initial delay can also be derived.

Comparisons between the model calculations and SPICE [8] simulations for wide ranges of device dimensions, input slope and loading capacitance are given for step and ramp inputs, respectively. Good agreement is obtained. The correlation between the initial delay and the propagation delay of CMOS inverters is observed in the characteristic waveform case [12]. The model developed in Section 2.2 is modified to account for the dependence of input slope on device parameters, in the characteristic waveform case, and is applied to investigate the initial delay or the propagation delay of CMOS inverters scaled down according to various scaling laws [13, 14]. The experimental results of initial delay for various device dimensions and input ramp rates are also shown.

2 Models

2.1 Step input

2.1.1 Initial fall delay: Consider a chain of p-well CMOS inverters as shown in Fig. 1. When a step input voltage $V_i(t)$ raises to V_{DD} instantly, the output voltage $V_o(t)$ first overshoots above V_{DD} and then decreases back to V_{DD} at $t = t_{df}$ (the initial fall delay). The input-voltage waveform and part of the output waveform are shown in Fig. 2*a* by solid lines in an enlarged scale; the waveforms in the undershoot case are also shown.

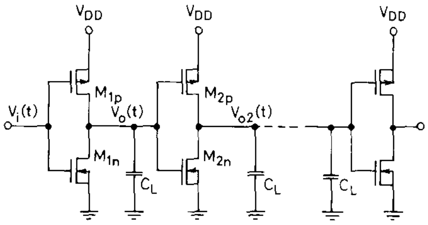

Fig. 1 *Chain of identical CMOS inverters*

To determine the operating regions of MOSFETs, the drain-source saturation voltage V_{dsat}, considering the velocity saturation effect, is calculated. The resultant V_{dsat} curve for the NMOS M_{1n} in the overshoot case is shown in Fig. 2*a*. It can be seen from Fig. 2*a* that M_{1n} is operating in the saturation region from $t = 0$ to $t = t_{df}$. During this period, however, the output node becomes the source node of the PMOS M_{1p} because its voltage is greater than V_{DD}, whereas the node connected to the power supply V_{DD} acts as the drain. This is opposite to the normal case and a positive substrate bias results, so that the magnitude of the PMOS threshold voltage is reduced [9]. M_{1p} will be turned on in the saturation region for a sufficiently large voltage overshoot. It is found, however, that the time interval when M_{1p} is turned on is rather short and the current generated by M_{1p} can be reasonably neglected, as compared with the transient current of the bulk-source junction capacitance C_{bsp}. M_{1p} is therefore assumed to be off during the overshoot period. The DC current of the bulk-source junction can also be neglected, as compared with its capacitance transient current in this short period.

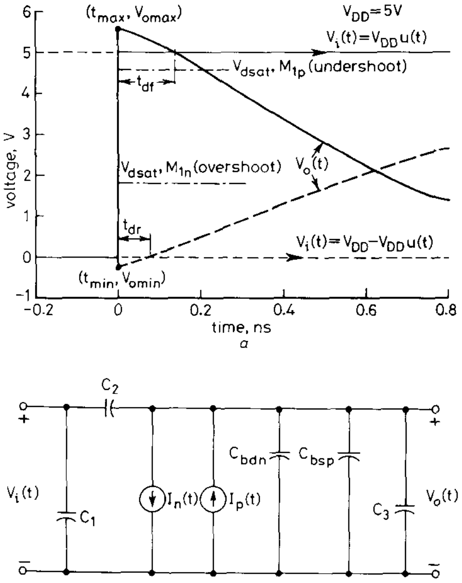

Fig. 2
a Schematic diagram of input and output voltage waveforms under step input
b Large-signal equivalent circuit of CMOS inverter chain

Since the voltage at the output node of the second stage remains unchanged during the overshoot period, the output node of the second stage can be effectively grounded. Thus, applying the MOSFET large-signal model to the four devices in the first two stages [2], the overall equivalent circuit during the initial fall delay can be obtained as shown in Fig. 2*b*. In Fig. 2*b*, the current $I_p(t)$ and the DC current of the source-substrate junction of M_{1p} are neglected, whereas $I_n(t)$ is the saturation current of M_{1n} and can be written as [8, 10]

$$\begin{aligned} I_n(t) = \beta_n\{&(V_{gsn} - V_{Tn} - \eta_n V_{dsatn}/2)V_{dsatn} - 2\gamma_{sn}/3 \\ &\times [(2\phi_{Fn} + V_{dsatn} - V_{bsn})^{3/2} - (2\phi_{Fn} - V_{bsn})^{3/2}]\} \\ &\times (1 + \lambda_n V_{dsn}) \qquad (1) \end{aligned}$$

where the drain-source saturation voltage V_{dsatn} can be solved from [15]

$$\begin{aligned} V_{maxn} = [\mu_n\{&(V_{gsn} - V_{Tn} - \eta_n V_{dsatn}/2)V_{dsatn} - 2\gamma_{sn}/3 \\ &\times [(2\phi_{Fn} + V_{dsatn} - V_{bsn})^{3/2} - (2\phi_{Fn} - V_{bsn})^{3/2}]\}] \\ &\div \{L_n[V_{gsn} - V_{Tn} - \eta_n V_{dsatn} - \gamma_{sn}(V_{dsatn} \\ &+ 2\phi_{Fn} - V_{bsn})^{1/2}]\} \qquad (2) \end{aligned}$$

Considering the velocity saturation effect, the channel-length modulation parameter λ_n in eqn. 1 can be expressed as [15]

$$\lambda_n = \{X_D[(X_D V_{maxn}/2/\mu_n)^2 + V_{dsn} - V_{dsatn}]^{0.5} - X_D^2 V_{maxn}/2/\mu_n\}/(L_n V_{dsn}) \qquad (3)$$

All other parameters involved in the $I_n(t)$ expression are further expressed in Table 1.

Table 1: Device parameters ued in MOSFET current equation

$$\beta_n = \mu_n \frac{W_n}{L_n} \frac{\varepsilon_{ox}}{T_{oxn}}$$

$$V_{Tn} = V_{Ton} - \gamma_n \sqrt{2\phi_{Fn}} + (\eta_n - 1)(2\phi_{Fn} - V_{bsn})$$

$$\eta_n = 1 + \frac{\pi}{4} \frac{\varepsilon_{si} T_{oxn}}{\varepsilon_{ox} W_n} \cdot DELTAn$$

$$\mu_n = \mu_{on} \left(\frac{\mu_{crit} \cdot \varepsilon_{si}}{\frac{\varepsilon_{ox}}{T_{oxn}} [V_{gsn} - V_{Tn} - \mu_{tra} \cdot \min(2\phi_{Fn}, V_{dsn})]} \right)^{\mu\,\mathrm{exp}}$$

$$\gamma_{sn} = \gamma_n (1 - a_s - a_d)$$

$$a_s = \frac{X_J}{2L_n} (\sqrt{1 + 2W_s/X_J} - 1)$$

$$a_d = \frac{X_J}{2L_n} (\sqrt{1 + 2W_D/X_J} - 1)$$

$$W_S = X_D \sqrt{2\phi_{Fn} - V_{bsn}}$$

$$W_D = X_D \sqrt{2\phi_{Fn} - V_{bdn}}$$

$$X_D = \sqrt{2\varepsilon_{si}/qN_{sub}}$$

The capacitances C_1, C_2 and C_3 in Fig. 2*b* can be expressed as [8, 11]

$$C_1 = C_{gson} + C_{gsop} + 2C_{on}/3 \qquad (4)$$

$$C_2 = C_{gdon} + C_{gsop} + 2C_{op}/3 \qquad (5)$$

$$C_3 = C_{gsop} + C_{gdop} + C_{gson} + C_{gdon} + C_{on} + C_{op} + C_L \qquad (6)$$

and the voltage-dependent p-n junction capacitances C_{bdn} and C_{bsp} can be written as [8]

$$C_{bdn}(C_{bsp}) = C_{Jn(p)} A_{sn(p)} (1 - V_{bdn(bsp)}/PB_{n(p)})^{-M_{Jn(p)}} + C_{JSWn(p)} P_{sn(p)} (1 - V_{bdn(bsp)}/PB_{n(p)})^{-M_{JSWn(p)}} \qquad (7)$$

In eqns. 4–6, the loading capacitance C_L and the gate-oxide capacitances C_{on} and C_{op} have larger effect on C_2 and C_3 than other capacitances.

To analytically solve the output voltage $V_o(t)$, necessary simplifications must be applied to linearise the current and capacitances. First, the channel-length modulation parameter λ_n and other V_{dsn}- and V_{gsn}-dependent parameters are all evaluated at $V_i = V_o = V_{DD}$. This leads to the linear dependence of $I_n(t)$ on V_{dsn}. Secondly, the capacitances C_{bdn} and C_{bsp} are assumed to be constant in solving $V_o(t)$. After the explicit expression of $V_o(t)$ is obtained, the voltage dependencies of C_{bdn} and C_{bsp} are incorporated into the expression of $V_0(t)$ to form a nonlinear equation. Iterations are then applied to get the final results.

Based upon the above assumptions, $V_0(t)$ can be solved as

$$V_o(t) = [(1 + C_2/C_2 + C_3 + C_{bdn} + C_{bsp}))V_{DD} + 1/\lambda_n]e^{-P_f t} - 1/\lambda_n \qquad (8)$$

where

$$P_f = \beta_n \lambda_n V_{DTn}^2/(C_2 + C_3 + C_{bdn} + C_{bsp}) \qquad (9a)$$

and

$$V_{DTn}^2 = (V_{DD} - V_{Tn} - \eta_n V_{dsatn}/2)V_{dsatn} - 2\gamma_{sn}/3 \times [(2\phi_{Fn} + V_{dsatn})^{3/2} - (2\phi_{Fn})^{3/2}] \qquad (9b)$$

From eqn. 8, the maximum output voltage V_{omax} occurs at $t = 0$ and is equal to

$$V_{omax} = V_{DD} + C_2 V_{DD}/(C_2 + C_3 + C_{bdn} + C_{bsp}) \qquad (10)$$

It can be seen from eqn. 10 that the voltage V_{omax} is larger than V_{DD} and the net voltage overshoot is just equal to the feedthrough voltage in the capacitance network of C_2, in series with the output capacitances C_3, C_{bdn}, and C_{bsp}. This is the cause of the initial delay.

At $t = 0$, the voltage V_{bdn} and V_{bsp} in eqn. 7 can be written as

$$V_{bdn} = -V_{omax} \qquad (11)$$

$$V_{bsp} = V_{omax} - V_{DD} \qquad (12)$$

Substituting eqns. 11 and 12 into eqn. 7 and then substituting eqn. 7 into eqn. 10, one can solve V_{omax}.

Define the initial fall delay t_{df} as the time interval from $t = 0$ to $t = t_{df}$ at which $V_0(t)$ is lowered back to V_{DD}. t_{df} can be obtained from eqn. 8 as

$$t_{df} = C_2 V_{DD}/[\beta_n \lambda_n V_{DTn}^2 (V_{omax} - V_{DD})] \times \ln[1 + (V_{omax} - V_{DD})/(V_{DD} + 1/\lambda_n)] \qquad (13)$$

where V_{omax} is given in eqn. 10. It is seen that the initial fall delay strongly depends on the device parameters as well as the power supply voltage. It should be noted that eqn. 13 is valid only for a finite λ_n which corresponds to the case of short-channel devices.

2.1.2 Initial rise delay: In the case of initial rise delay, the input is a falling step input. The input- and output-voltage waveforms are also shown in Fig. 2*a* by dashed lines. According to the modelling method in Section 2.1.1, the voltage magnitude V_{omin} at the minimum point can be similarly derived and expressed as

$$V_{omin} = C_2 V_{DD}/(C_2 + C_3 + C_{bsn} + C_{bdp}) \qquad (14)$$

The initial rise delay t_{dr}, defined as the time interval from $t = 0$ to $t = t_{dr}$ at which $V_0(t)$ is raised back to 0 V, can be obtained

$$t_{dr} = C_2 V_{DD}/(\beta_p \lambda_p V_{DTp}^2 V_{omin}) \times \ln[1 + V_{omin}/(V_{DD} + 1/\lambda_p)] \qquad (15)$$

where β_p, λ_p and V_{DTp}^2 are all linearised by the same method in Section 2.1.1.

2.2 Ramp input

2.2.1 Initial fall delay: Fig. 3 shows the rising (falling) ramp input $V_i(t)$ and the corresponding output waveform $V_o(t)$ in solid (dashed) lines at the input and output nodes of the first stage in Fig. 1. In the case of rising ramp input, the displacement current flowing from the input node to output node is initially larger than the drain linear current $I_p(t)$ of M_{1p}, which has the output node as its source node. Therefore, a net charging current is presented at the output node so that $V_o(t)$ increases above V_{DD}. As $V_i(t)$ increases to turn on the NMOS M_{1n} in the saturation region, the increase of total drain current ($I_p(t)$

$+ I_n(t))$ flowing out of the output node tends to decrease the net current charging to the output node. This reduces the rate of increase of $V_o(t)$. When $V_o(t)$ reaches the maximum value V_{omax} at $t = t_{max}$, the net current charging to the output node is reduced to zero and $V_o(t)$ begins to fall to ground after t_{max}.

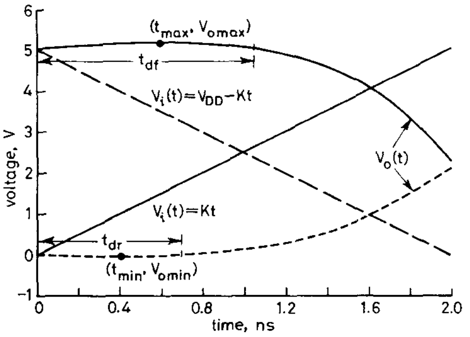

Fig. 3 *Schematic diagram of input and output voltage waveforms under ramp input*

$V_{DD} = 5$ V
$K = 2.5 \times 10^9$

Refering to the circuit in Fig. 2*b*, we have

$$C_2 \frac{\delta(V_i - V_o)}{\delta t} = I_n(t) - I_p(t) + (C_3 + C_{bdn} + C_{bsp}) \frac{\delta V_o}{\delta t} \tag{16}$$

where $I_n(t)$ is given in eqn. 1 and $I_p(t)$ is the drain linear current of M_{1p} and can be expressed as

$$I_p(t) = -\beta_p\{(V_{sgp} - V_{Tp} - \eta_p V_{sdp}/2)V_{sdp} - 2\gamma_{sp}/3 \times [(2\phi_{Fp} + V_{sdp} - V_{sbp})^{3/2} - (2\phi_{Fp} - V_{sbp})^{3/2}]\} \tag{17}$$

Note that the capacitance C_2 is equal to $(C_{gdon} + C_{gsop} + C_{op}/2)$ and the capacitance $C_{bdn}(C_{bsp})$ is evaluated at $V_{bdn}(V_{bsp}) = V_{DD}(0)$.

At $t = t_{max}$, that the charging current at the output node is zero implies

$$\left.\frac{\delta V_o}{\delta t}\right|_{t=t_{max}} = 0 \tag{18}$$

With the ramp rate K, the input voltage is written as

$$V_i(t) = Kt \tag{19}$$

Substituting eqns. 1, 17, 18 and 19 into eqn. 16, we have

$$\begin{aligned} C_2 K = {} & \beta_n\{(Kt_{max} - V_{Tn} - \eta_n V_{dsatn}/2)V_{dsatn} - 2\gamma_{sn}/3 \\ & \times [(2\phi_{Fn} + V_{dsatn})^{3/2} - (2\phi_{Fn})^{3/2}]\}(1 + \lambda_n V_{omax}) \\ & + \beta_p[(1 - \eta_p/2)V_{omax}^2 + (V_{bp} - Kt_{max})V_{omax} \\ & + V_{DD} Kt_{max} + V_{cp}^2] \end{aligned} \tag{20}$$

where

$$V_{bp} = (\eta_p - 1)V_{DD} - V_{Tp} - 2\gamma_{sp}3 \cdot (2\phi_{Fn})^{1/2} \tag{21a}$$

and

$$V_{cp}^2 = V_{DD} V_{Tp} - \eta_p V_{DD}^2/2 - 2\gamma_{sp}/3[(2\phi_{Fn})^{3/2} - (2\phi_{Fp} + V_{DD})(2\phi_{Fp})^{1/2}] \tag{21b}$$

In eqn. 20, V_{dsatn} and λ_n are functions of V_{omax} and t_{max} and can be calculated from eqns. 2 and 3, respectively.

Since the currents must be continuous, eqn. 16 can be differentiated as

$$C_2 \frac{\delta^2(V_i - V_o)}{\delta t^2} = \frac{\delta I_n(t)}{\delta t} - \frac{\delta I_p(t)}{\delta t} + (C_3 + C_{bdn} + C_{bsp}) \frac{\delta^2 V_o}{\delta t^2} \tag{22}$$

From eqns. 1, 2, 17, 18 and 19, eqn. 22 can be further expressed at $t = t_{max}$ as

$$\begin{aligned} & \beta_n[(K - \eta_n V'_{dsatn}/2)V_{dsatn} \\ & \quad + (Kt_{max} - V_{Tn} - \eta_n V_{dsatn}/2)V'_{dsatn} \\ & \quad - \gamma_{sn}(2\phi_{Fn} + V_{dsatn})^{1/2} V'_{dsatn}](1 + \lambda_n V_{omax}) \\ & \quad + \beta_p K(V_{DD} - V_{omax} + (C_2 + C_3 + C_{bdn} + C_{bsp})V''_o \\ & = 0 \end{aligned} \tag{23}$$

where

$$\begin{aligned} V'_{dsatn} = {} & K(V_{maxn} L_n - \mu_n V_{dsatn})/\{V_{maxn} L_n \\ & \times [\eta_n + \gamma_{sn}/2 \cdot (V_{dsatn} + 2\phi_{Fn})^{-1/2}] \\ & + \mu_n[Kt_{max} - V_{Tn} - \eta_n V_{dsatn} \\ & - \gamma_{sn}(2\phi_{Fn} + V_{dsatn})^{1/2}]\} \end{aligned} \tag{24a}$$

and

$$V''_o = \left.\frac{\delta^2 V_o}{\delta t^2}\right|_{t=t_{max}} \cong 2(V_{DD} - V_{omax})/t_{max}^2 \tag{24b}$$

The detailed derivations of eqn. 24*b* are shown in the Appendix. From eqns. 20 and 23, V_{omax} and t_{max} can be obtained. After solving V_{omax} and t_{max}, the initial fall delay t_{df} can be calculated as follows: since the output voltage $V_o(t)$ is equal to V_{DD} at $t = t_{df}$, the drain linear current $I_p(t)$ is equal to zero and eqn. 16 can be simply written as

$$KC_2 = \beta_n V_{dsatn}(Kt_{df} - V_{TEn} - \eta_n V_{dsatn}/2) \times (1 + \lambda_n V_{DD}) + (C_2 + C_3 + C_{bdn} + C_{bsp})V'_o \tag{25}$$

where

$$V_{TEn} = V_{Tn} + 2\gamma_{sn}/(3V_{dsatn}) \times [(2\phi_{Fn} + V_{dsatn})^{3/2} - (2\phi_{Fn})^{3/2}] \tag{26a}$$

and

$$V'_o = \left.\frac{\delta V_o}{\delta t}\right|_{t=t_{df}} \cong (V_{omax} - V_{DD})/(t_{max} - t_{df}) \tag{26b}$$

t_{df} can then be obtained from eqn. 25 with V_{omax} and t_{max} known from eqns. 20 and 23. Note that, in the above calculation, the parameters μ_n and λ_n are evaluated at $V_{dsn} = V_{DD}$.

Based upon eqns. 20, 23 and 25, the dependence of the initial fall delay on the input ramp rate, loading capacitance and device parameters can be calculated.

2.2.2 *Initial rise delay:*

In the case of initial rise delay, the analysis proceeds in a manner similar to that in Section 2.2.1. The input voltage is a falling-ramp waveform

$$V_i(t) = V_{DD} - Kt \tag{27}$$

t_{dr} can be obtained after solving V_{omin} and t_{min}, similar to the case of t_{df}.

3 Comparison to SPICE simulation

3.1 Step input

Fig. 4 shows the calculated initial fall delay t_{df} together with SPICE simulation results. The effective channel length for both NMOS and PMOS varies from 3 μm to 1 μm. For simplicity, the device parameters for $L_{eff} =$ 1 μm are assumed to be the same as those for $L_{eff} =$ 2 μm. The ratio of the aspect ratio (width/length) of PMOS to that of NMOS is equal to unity or the ratio of their carrier mobilities. To calculate the source/drain areas, the dimension W_{ex} of the source/drain regions in the channel-length direction is assumed to be 8.5 μm for both PMOS and NMOS. W_{ex} is proportionally scaled for devices with smaller effective channel lengths. In all these cases, a general agreement between model calculations and SPICE simulations is obtained.

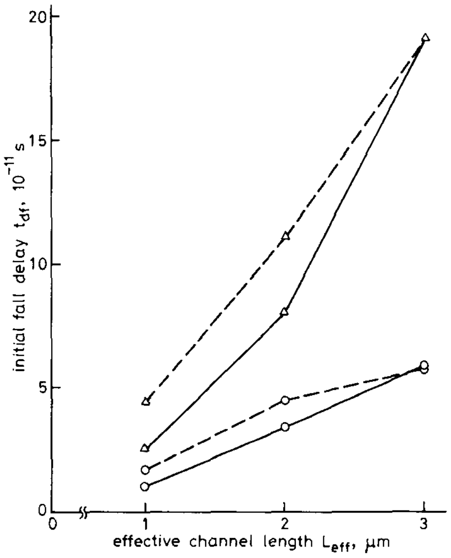

Fig. 4 *Initial fall delay for various channel lengths under step input*

——— Simulation
– – – – Calculation
○○○ $W_p = W_n = 32$ μm
△△△ $W_p = (\mu_n/\mu_p)W_n$
$L_n = L_p = L_{eff}$
$C_L = 0$
For $W_p = (\mu_n/\mu_p)W_n$

$$W_n = \begin{cases} 3.5\ \mu\text{m},\ L_{eff} = 3\ \mu\text{m} \\ 2.33\ \mu\text{m},\ L_{eff} = 2\ \mu\text{m} \\ 1.17\ \mu\text{m},\ L_{eff} = 1\ \mu\text{m} \end{cases}$$

When the device dimensions are scaled down, both C_2 and C_3 of NMOS and PMOS are decreased, while λ_n is increased. The resultant t_{df} is decreased as predicted by eqn. 13 and is shown in Fig. 4. In general, a large t_{df} leads to large gate delay in CMOS logic gates.

Fig. 5 shows both calculated and SPICE-simulated initial rise delay t_{dr}. A satisfactory agreement is achieved. As shown in Fig. 5, t_{dr} is larger for $L_{eff} = 2$ μm than for $L_{eff} = 3$ μm. This is a special case due to the inherent transconductance degradation of PMOS in the adopted 2 μm CMOS process. Generally, t_{dr} decreases with a decrease in the channel length, as one compares the case of $L_{eff} = 2$ μm to that of $L_{eff} = 1$ μm.

It can be seen from Figs. 4 and 5 that the initial delays vary with the channel-width ratio W_p/W_n. Fig. 6 shows the variations of the initial fall delay, the initial rise delay and total initial delay ($t_{df} + t_{dr}$) with channel-width ratio W_p/W_n for $W_n = 2$ μm. It is seen from Fig. 6 that the minimum initial delay of a CMOS inverter can be achieved with a suitable value of W_p/W_n. According to

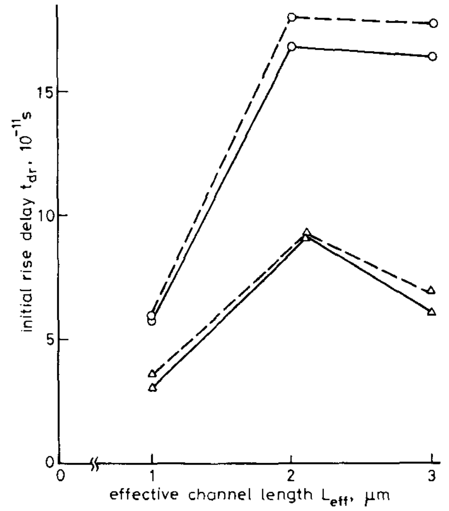

Fig. 5 *Initial rise delay for various channel lengths under step input*

——— Simulation
– – – Calculation
○○○ $W_p = W_n = 32$ μm
△△△ $W_p = (\mu_n/\mu_p)W_n$
$L_n = L_p = L_{eff}$
$C_L = 0$
For $W_p = (\mu_n/\mu_p)W_n$

$$W_n = \begin{cases} 3.5\ \mu\text{m},\ L_{eff} = 3\ \mu\text{m} \\ 2.33\ \mu\text{m},\ L_{eff} = 2\ \mu\text{m} \\ 1.17\ \mu\text{m},\ L_{eff} = 1\ \mu\text{m} \end{cases}$$

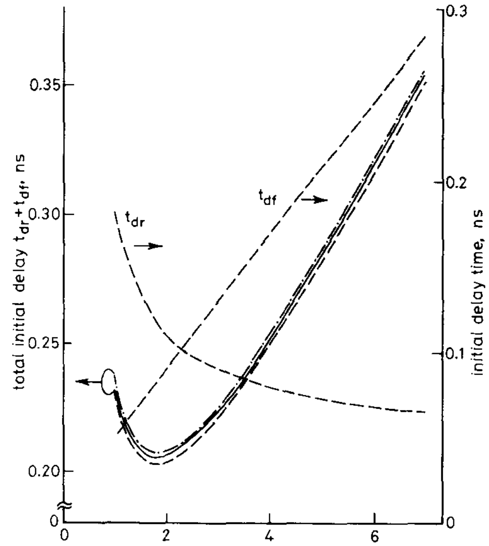

Fig. 6 *Variations of initial delays with channel-width ratio W_p/W_n for different loading capacitance under step input*

$$C_L = \begin{cases} \text{- - - -}\quad 0\ \text{P} \\ \text{———}\quad 0.2\ \text{P} \\ \text{—·—·}\quad 0.8\ \text{P} \end{cases}$$

$L_n = L_p = 2$ μm
$W_n = 2$ μm

eqn. 13 and 15, if the variations of V_{omax} and V_{omin} are negligible, the optimal width ratio W_p/W_n, which results in the minimum total initial delay, can be approximately obtained as

$$(W_p/W_n)_{opt} = \sqrt{\frac{\mu_n T_{oxp} \lambda_n V_{DTn}^2 L_p \ln [1 + V_{omin}/(V_{DD} + 1/\lambda_p)]}{\mu_p T_{oxn} \lambda_p V_{DTp}^2 L_n \ln [1 + (V_{omax} - V_{DD})/(V_{DD} + 1/\lambda_n)]}} \tag{28}$$

As can be seen from eqn. 28, the optimal ratio W_p/W_n is proportional to the square root of the mobility ratio μ_n/μ_p. This is consistent with the results of previous work [3, 7].

3.2 Ramp input

Fig. 7*a* shows the calculated and SPICE-simulated dependence of t_{df} and t_{dr} on W_p/W_n with the ramp rate K as a parameter. Good agreement is obtained. As can be seen from Fig. 7*a*, t_{df} increases, whereas t_{dr} decreases, with the increase of W_p/W_n. The negative dependence of t_{dr} on W_p/W_n results from the increase of W_p increasing the the drain current of M_{1p} charging to the output node. Therefore, for a fixed value of K, there exists an optimal W_p/W_n such that the total initial delay $(t_{df} + t_{dr})$ achieves a minimum value, similar to the case of step input. The resultant $(t_{df} + t_{dr})$ is shown in Fig. 7*b* for different values of K. It is seen from Fig. 7*b* that the optimal W_p/W_n, resulting in the minimum $(t_{df} + t_{dr})$, is equal to 1.75 for all K, which is close to the value 1.78 calculated from eqn. 28. This reveals that the optimal channel-width ratio is independent of input excitations and is determined only by the device parameters as approximately expressed in eqn. 28.

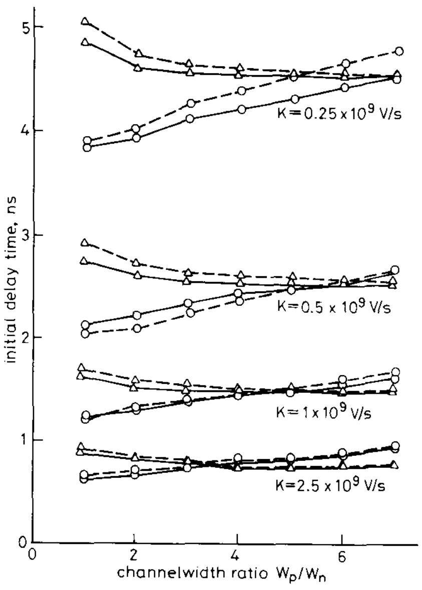

Fig. 7A *Variations of initial fall and initial rise delays*

$L_n = L_p = 2\ \mu m$
$W_n = 2\ \mu m$
$C_L = 0$

Simulation { —○—○ t_{df}
—△-—△ t_{dr}

Calculation { - ○--○ t_{df}
-△--△ t_{dr}

The effect of loading capacitance C_L on the initial delay is shown in Fig. 8, where SPICE-simulation results are also given for comparison. As can be seen from Fig. 8,

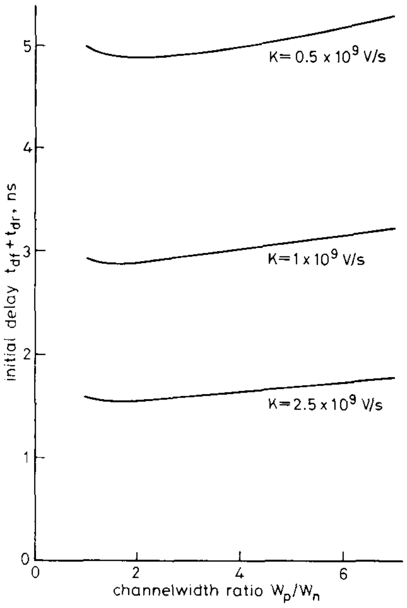

Fig. 7B *Variations of initial delay $(t_{df} + t_{dr})$ with channel-width ratio W_p/W_n for different ramp rate under ramp input*

$L_n = L_p = 2\ \mu m$
$W_n = 2\ \mu m$
$C_L = 0$

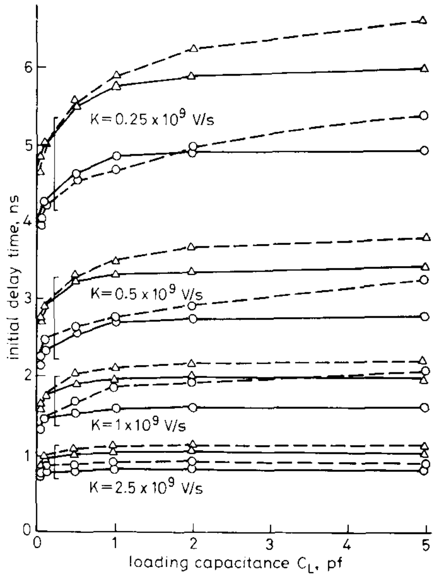

Fig. 8 *Dependence of initial delays on loading capacitance for different ramp rate under ramp input*

Simulation { - ○—○ t_{df}
—△ -△ t_{dr}

Calculation { --○ -○ t_{df}
- -△--△ t_{dr}

$L_n = L_p = 2\ \mu m$
$W_n = 2\ \mu m$
$W_p = 4\ \mu m$

both t_{df} and t_{dr} increase with the increase of loading capacitance. However, the rate of increase of t_{df} or t_{dr} is not linearly proportional to the loading capacitance. As the loading capacitance increases, the net current charging to the output node is reduced, due to the increased transient current across C_L, so that the output voltage overshoot V_{omax} is decreased. However, the decrease of V_{omax} in turn decreases the device current $I_p(t)$ flowing out of the output node. Thus the resultant effects lead to a small increase of the initial delays when increasing the loading capacitance.

It can be seen from Figs. 7 and 8 that the dominant factor in affecting the initial delays is the input ramp rate K. Fig. 9 shows the calculated and simulated variations of t_{df} and t_{dr} with the value V_{DD}/K for different values of K. Also shown in Fig. 9 are the corresponding values of input voltage $V_{ir}(V_{if})$ calculated from eqn. 19 (eqn. 27) at $t = t_{df}(t_{dr})$. The initial delays increase drastically and tend to be linearly proportional to V_{DD}/K as K decreases. However, the linear dependence of t_{df} and t_{dr} on V_{DD}/K occurs only for small values of K (slowly ramped input), as shown in Fig. 9, When the ramp rate becomes large, the feedthrough currents from input node to the output node are so large that more current $I_n(t)$ or $I_p(t)$ is needed to draw the output node to its normal state. Thus, the input must take more time to increase to a sufficient value above the threshold voltage of M_{1n} or M_{1p} to increase $I_n(t)$ or $I_p(t)$. This can be seen from the value of $V_{ir}(V_{if})$ in Fig. 9, where $V_{ir}(V_{DD} - V_{if})$ increases monotincially beyond the threshold voltage $V_{Tn}(V_{Tp})$ of $M_{1n}(M_{1p})$ for a large ramp rate. Therefore, only when the rising or falling time of the input voltage is very long, can the initial delays be calculated as the time at which the input voltage reaches the threshold voltage of NMOS or PMOS. Otherwise, the capacitance feedthrough effects should be taken into account in calculating the initial delay time as considered in Section 3.1.

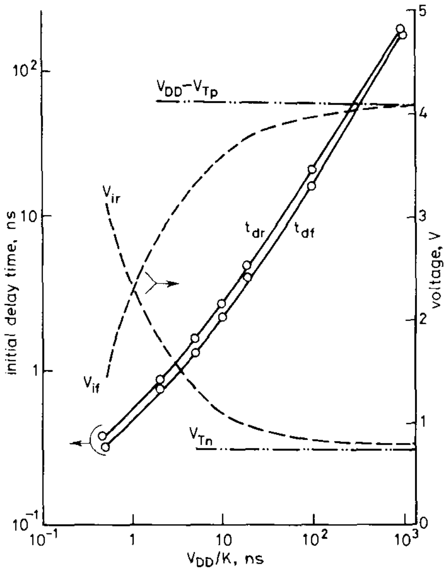

Fig. 9 *Calculated variations of initial fall and initial rise delays and corresponding input voltages with different ramp rate under ramp input*

○○○ simulation
—— – calculation
$L_n = L_p = 2\ \mu m$
$W_n = 2\ \mu m$
$W_p = 4\ \mu m$
$C_L = 0$
$V_{DD} = 5$ V
$V_{ir} = Kt_{df}$
$V_{if} = V_{DD} - Kt_{dr}$

4 Propagation delay

In the delay time evaluation, the commonly used parameter is the propagation delay or pair delay of a logic gate. In the characteristic waveforms [12], as shown in Fig. 10, the successive waveforms at the output nodes of intermediate stages in a chain of identical CMOS inverters are plotted. The propagation delay t_p is defined as the time interval between the successive rising or falling waveforms at the voltage level of half-signal swings ($V_{DD}/2$). Owing to the capacitive feedthrough effect, the voltage overshoot and undershoot can be clearly observed. In the characteristic waveforms, the initial fall delay t_{df} is defined as the time interval from the point of $V_i(t)$ at which $V_i(t) = 0$ V (point A) to the point of $V_{o1}(t)$ at which $V_{o1}(t) = 5$ V (point B). Whereas, the initial rise delay t_{dr} is defined as the time interval from the point of $V_{o1}(t)$ at which $V_{o1}(t) = 5$ V (point B) to the point of $V_{o2}(t)$ at which $V_{o2}(t) = 0$ V (point C).

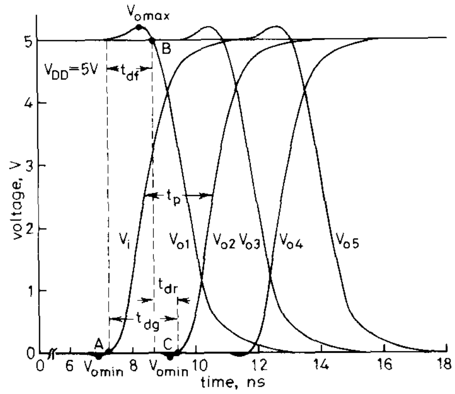

Fig. 10 *Typical characteristic waveforms of chain of identical CMOS inverters together with the definitions of propagation delay and gate initial delay*

With the specified initial delays, the sum of the initial fall delay and the initial rise delay, defined as the gate initial delay t_{dg}, is just equal to the propagation delay t_p, as verified in Fig. 11, where the propagation delay against the gate initial delay is plotted from SPICE results for several chains of identical CMOS inverters. It is not surprising, as shown in Fig. 11, that the propagation delay is equal to the gate initial delay. When $V_i(t)$ in Fig. 10 increases from 0 V, an initial fall delay is taken before the output voltage $V_{o1}(t)$ decreases below V_{DD}. Meanwhile, as $V_{o1}(t)$ decreases from V_{DD} a second time, the initial rise delay is spent before $V_{o2}(t)$ increases above 0 V. Since $V_i(t)$ and $V_{o2}(t)$ are duplicate waveforms with equal rise times, the propagation delay from $V_i(t)$ to $V_{o2}(t)$ therefore results from the sum of the initial fall and the initial rise delays. The smaller the gate initial delay, the smaller the propagation delay will be.

Since the model developed in Section 2.2 considers the small-geometry effects in short-channel devices, it is used to calculate the delay performance of a chain of scaled CMOS inverters. In the calculations, a set of 3 μm device parameters in Table 2, for both NMOS and PMOS, are scaled simultaneously in accordance with the three scaling laws: the constant-electric-field (CE), the constant-voltage (CV) and the quasiconstant-voltage (QCV) laws [13, 14] as shown in Table 3, where s is the scaling factor. The zero-biased threshold voltages for both NMOS and PMOS are scaled in the same way as that for the power supply voltage.

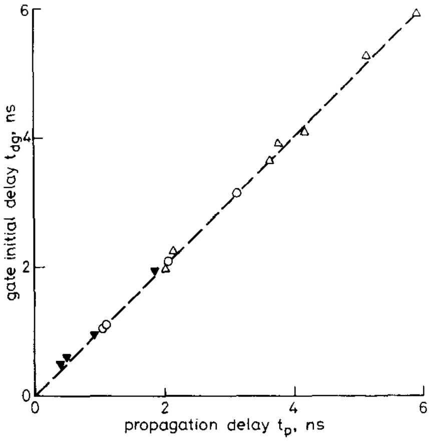

Fig. 11 *Correlation between gate initial delay and propagation delay obtained from SPICE simulations for various device dimensions*

△ 3 μm, ○ 2 μm, ▼ 1 μm $\}= L_n = L_p$

– – – – $t_p = t_{dg}$

Table 2: Device parameters to be scaled according to scaling laws in Table 3

Parameter	NMOS	PMOS
L, μm	3.0	3.0
W, μm	3.0	6.0
V_{To}, V	0.725	0.725
T_{ox}, m	7.01E-8	6.48E-8
PB, V	0.775	0.85
N_{sub}, cm^{-3}	2.0E16	1.75E15
V_{max}, m/s	8.7E4	6.4E4
X_J, μm	0.5	0.5
$DELTA$	0.997	0.
μ_o, m^2/Vs	820E-4	250E-4
μ_{exp}	0.113	0.295
μ_{crit}, V/m	6.377E6	6.024E6
μ_{tra}	0.2	0.3
M_J	0.543	0.515
M_{JSW}	0.34	0.341
C_J, f/m^2	3.8E-4	1.2E-4
C_{JSW}, f/m	3.6E-10	3.6E-10
C_{gso}, f/m	1.23E-10	1.33E-10
C_{gdo}, f/m	1.23E-10	1.33E-10
W_{ex}, μm	8.5	8.5

$V_{DD} = 5$ V

Table 3: Three scaling laws for scaling of device parameters

Scaling law	Constant field	Constant voltage	Quasi constant voltage
Dimension	s^{-1}	s^{-1}	s^{-1}
Voltage	s^{-1}	I	$s^{-0.5}$
Oxide thickness	s^{-1}	$s^{-0.5}$	s^{-1}
Dopant	s	s	s

Two cases, the constant ramp rate and the scaled ramp rate, are considered. In the case of constant ramp rate, the initial delay of the first stage in Fig. 1 is calculated under a fixed ramp-input excitation. In the case of scaled ramp rate, the input waveform in the characteristic waveforms case is simulated by a ramp waveform, with an appropriate ramp rate [7], to calculate the gate initial delay or propagation delay. In this case, the ramp rate of the input waveform of the present stage is also affected by the parameters of the previous stage. According to the expression of eqn. 9*a*, the ramp rate of the falling waveform can be approximately expressed as

$$K = (\beta_n V_{DTn}^2 \lambda_n V_{DD})/(C_2 + C_3 + C_{bdn} + C_{bsp}) \qquad (29a)$$

whereas that for the rising waveform is

$$K = (\beta_p V_{DTp}^2 \lambda_p V_{DD})/(C_2 + C_3 + C_{bsn} + C_{bdp}) \qquad (29b)$$

For the adopted parameters in Table 2, the calculated t_{dr} (t_{df}) by using eqn. 29*a* (eqn. 29*b*) is 0.8038 ns (1.103 ns), which agrees with the value 0.8027 ns (1.1345 ns) obtained from the characteristic waveform of SPICE simulations. This shows that the initial delays of characteristic waveforms can be calculated by simulating the input waveform with an appropriate ramp waveform.

Applying the model in Section 2.2, Figs. 12*a* and 12*b* show the calculated dependence of the initial delay ($t_{df} + t_{dr}$) and gate initial delay t_{dg} on the scaling factor s for constant ramp rate and scaled ramp rate, respectively. As can be seen from Fig. 12, the initial delay decreases with an increasing scaling factor s for all three scaling laws, Moreover, the linear dependence of the delays ($t_{df} + t_{dr}$) and t_{dg} occurs for the CE law, which agrees with the results predicted by the conventional first-order analysis.

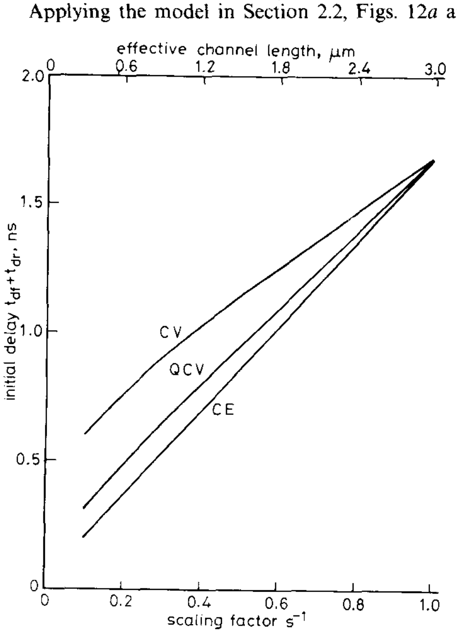

Fig. 12A *Predictions of initial delay ($t_{df} + t_{dr}$) under constant ramp rate*
$K = 2.5 \times 10^9$ V/s

When the input is driven by a waveform with a constant ramp rate, the scaling of devices with the CE law leads to the greatest improvement in the delay of the first stage than that with the other two laws, as can be seen from Fig. 12*a*. Furthermore, because the large driving capability of CMOS devices is offset by the effects of a large voltage swing in the CV law, the CV law leads to the least delay improvement for constant ramp rate.

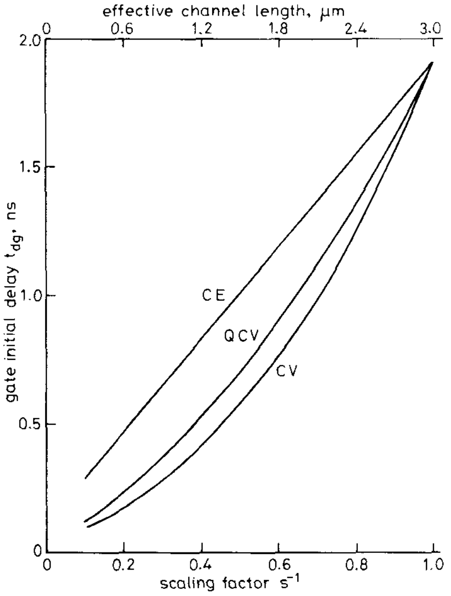

Fig. 12B *Gate initial delay under scaled ramp rate in chain of CMOS inverters as function of scaling factor s for CE, CV and QCV laws*

$t_p = t_{dq}$

In the case of the characteristic waveform, however, the ramp rate of the input waveform is also affected, through eqns. 29*a* and 29*b*, by the scaling laws applied to the previous stage. Thus, in contrast to those in the constant ramp rate case, the reduction of gate initial delay for the QCV law is larger than that for the CE law but smaller than that for the CV law, as shown in Fig. 12*b*. This reveals that as one tries to improve the speed of internal circuits by the CV law, one will obtain less delay improvement in the input stage by the CV law than that by the other two laws. Therefore, the most effective manner in which to apply the scaling laws for delay improvement is to apply the CV law to the internal circuits, whereas the CE law is applied to the input stage of internal circuits.

5 Experiment

Three chains of CMOS inverters were fabricated with 3 μm p-well technology. The channel lengths of both NMOS and PMOS are 3 μm, 5 μm and 8 μm for the three chains. The channel width of the NMOS is 4 μm, 10 μm and 16 μm and that of the corresponding PMOS is 8 μm, 20 μm or 16 μm for the three chains. Each output node of the inverter chains is connected, through a two-stage buffer, to an output pad. The dimension of the first stage of the output buffer is the same as that of the corresponding inverter chain. With a 5 V power supply, Fig. 13 shows the measured waveform at the output node of the first stage. It is clearly seen that the voltage overshoot and undershoot indeed occur during the transitions of input waveform.

To investigate the initial delay, a ramp input with rising or falling ramp rate equal to $V_{DD}/10$ ns, $V_{DD}/20$ ns,

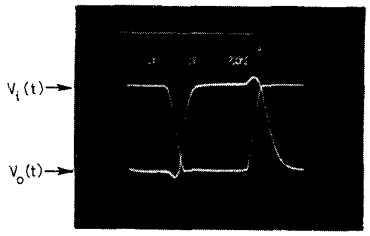

Fig. 13 *Experimental output waveform at output node of first stage of chain of CMOS inverters under ramp input*

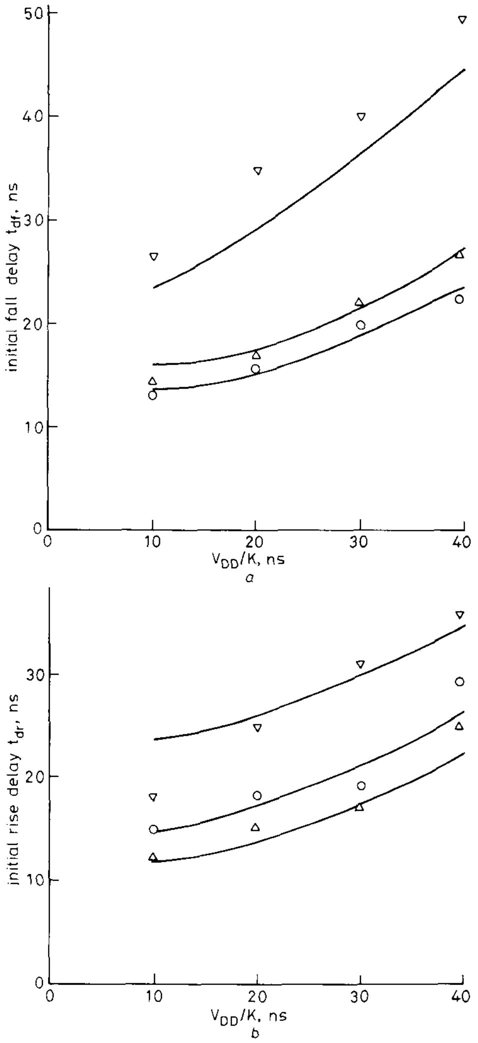

Fig. 14 *Comparisons between theoretical and experimental results*

a initial fall delay
b initial rise delay for various input ramp rates
—— calculation
○ $(W/L)_p = 8/3.5$ $(W/L)_n = 4/3.5$ (μm/μm)
△ $(W/L)_p = 20/5$ $(W/L)_n = 10/5$ (μm/μm)
▽ $(W/L)_p = 16/8$ $(W/L)_n = 16/8$ (μm/μm)
} experiment
$V_{DD} = 5$ V

$V_{DD}/30$ ns or $V_{DD}/40$ ns was applied to the input node of the first stage, separately. The voltage waveform at the output node of the first stage was then picked up by the FET probe with 2 pf input capacitance. The parasitic capacitance between the input and output terminals due to wirings was estimated to be 0.08 pf. Taking all the factors into considerations, Figs. 14*a* and 14*b* show the calculated and experimental results of the initial fall and initial rise delays under various ramp inputs, respectively. Satisfactory agreement is obtained between the theory and experiment.

6 Conclusion

The initial delays in a CMOS inverter under step- and ramp-input excitations are characterised in detail. It is found that both the initial rise and initial fall delays determine the propagation delay of CMOS logic gates. Thus, the initial delay actually determines the speed of a logic gate. From this point of view, the initial delays are modelled for inverters with different dimensions. Optimal device dimension that results in minimum initial delay is obtained. The results are compared to those from SPICE simulations and good agreement between theoretical and SPICE-simulated results is obtained. It is shown that the capacitive feedthrough effect and thus the input slope are important factors in determing the initial delay.

Based on the model, the delay time for scaled-down CMOS inverters is calculated. As the channel length is scaled, the delay is decreased as expected. Moreover, the calculations show that the constant-voltage-scaling law is the best choice of scaling laws in reducing the propagation delay of characteristic waveforms. However, for the delay response of the input stage, the constant-electric-field law is found to be the most effective law in reducing the initial delays.

Although the analysis concentrates on CMOS inverters only, other logic gates such as NAND, NOR, and transmission gate also have initial delays and similar modelling method can be applied accordingly.

7 References

1 MDINDL, J.D.: 'Theoretical, practical and analogical limits in ULSI', *Int. Electron Devices Meet. Tech. Dig.*, 1983, pp. 8–13

2 WU, C.-Y., HWANG, J.-S., CHANG, C., and CHANG, C.-C.: 'An efficient timing model for combinational logic gates', *IEEE Trans.*, 1985, **CAD-4**, pp. 636–650

3 KUNAMA, A.: 'CMOS circuit optimization', *Solid-State Electron.*, 1983, **26**, pp. 47–58

4 TOKUDA, T., OKAZAKI, K., SAKASHITA, K., OHKURA, I., and ENOMOTO, T.: 'Delay-time modelling for ED MOS logic LSI', *IEEE Trans.*, 1983, **CAD-2**, pp. 129–134

5 SIMMONS, J.G., and TAYLOR, G.W.: 'An analytic treatment of the performance of submicrometer FET logic', *IEEE J. Solid-State Circuits*, 1985, **SC-20**, pp. 1242–1251

6 AUVERGNE, D., CAMBON, G., DESCHACHT, D., ROBERT, M., SAGNES, G., and TEMPIER, V.: 'Delay-time evaluation in ED MOS logic LSI', *IEEE J. Solid-State Circuits*, 1986, **SC-21**, pp. 337–343

7 HEDENSTIERNA, N., and JEPPSON, K.O.: 'CMOS circuit speed and buffer optimization', *IEEE Trans.*, 1987, **CAD-6**, pp. 270–281

8 NAGEL, L.W.: 'SPICE 2: a computer program to simulate semiconductor circuits'. University of California, Berkeley, CA, USA, 1975

9 RICHMAN, P.: 'MOS field-effect transistor and integrated circuits' (John Wiley Inc., New York, 1973)

10 YAU, L.D.: 'A simple theory to predict the threshold voltage of short-channel IGFET's', *Solid-State Electron.*, 1974, **17**, pp. 1059–1063

11 ELMASRY, M.I.: 'Digital MOS integrated circuits: A Tutorial', *in* 'Digital MOS Integrated Circuits' (IEEE Press, 1981), pp. 4–27

12 BURNS, J.R.: 'Switching response of complementary-symmetry MOS transistor logic circuit', *RCA Rev.*, 1964, pp. 627–661

13 DANNARD, R.H., GAENSSLEN, F.H., YU, H.-N., RIDEOUT, V.L., BASSOUS, E., and LABLANE, A.R.: 'Design of ion-implanted MOSFET's with very small physical dimensions', *IEEE J. Solid-State Circuits*, 1974, **SC-9**, pp. 256–267

14 VLSI Lab., T.I. Inc.: 'Technology and design challenges of MOS VLSI', *IEEE J. Solid-State Circuits*, 1982, **SC-17**, pp. 442–448

15 VLADIMERSCU, A., and LIU, S.: 'The simulation of MOS integrated circuits using SPICE 2'. UCB/ERL M8017, Electronics Research Lab., University of California, Berkeley, CA, USA, 1980

8 Appendix

The Taylor's expansion of $V_o(t)$ around $t = t_{max}$ can be written as

$$V_o(t) = V_{omax} + (t - t_{max})V'_o(t_{max}) + (t - t_{max})^2 V''_o(t_{max})/2 + H(t - t_{max}) \quad (30)$$

where $H(t - t_{max})$ are those of higher order terms. The boundary condition of $V_o(t)$ is given by

$$V_o(0) = V_{DD} \quad (31)$$

Substituting eqns. 31 and 18 into eqn. 30, the second derivative of $V_o(t)$ at $t = t_{max}$ can be written as

$$V''_o(t_{max}) = 2[V_{DD} - V_{omax} - H(t_{max})]/t^2_{max} \quad (32)$$

By neglecting the higher order terms in eqn. 32, eqn. 24*b* can be obtained.